A Multi-subject Temporal-spatial Hyper-alignment Method for EEG-based Neural Entrainment to Speech

Di Zhou*, Gaoyan Zhang†, Jianwu Dang*†, Shuang Wu† and Zhuo Zhang†
* Japan Advanced Institute of Science and Technology, Japan
E-mail: zhoudi@jaist.ac.jp, jdang@jaist.ac.jp
† College of Intelligence and Computing, Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin University, China
E-mail: zhanggaoyan@tju.edu.cn

***Abstract*—The low signal-to-noise ratio (SNR) of a neural recording is typically improved by averaging the neural response over repeated trials. However, it is not applicable when studying neural entrainment to speech stimuli, in which stimuli are presented only once. Alternatively, multiple subjects' neural responses to the same stimuli can be averaged to decrease unexpected noises caused by breathing, lack of attentiveness etc., excluding those caused by heartbeats and blinking during long auditory tasks. However, individual differences such as varying latency times of the neural response to the stimulus and electrode positioning in the setup reduce the effectiveness of this method. To eliminate individual differences, we first estimated the importance (weight) of each electrode using a spatial filter and maximized the SNR by adjusting the latency of the neural response. Then we extracted the common response for all subjects and constructed neural entrainment models accordingly. The correlations between the predicted and actual neural responses obtained in this study were much higher than that in other methods in the forward neural encoding process. In the decoding process, the correlation between the reconstructed speech envelope and the original also increased significantly in both the delta and theta bands compared with previous studies.**

I. Introduction

Speech perception, which links auditory and cognitive processes, is the acquisition of communicative information from speech sounds [1]. One of the main objectives of auditory neuroscience research is to investigate how ubiquitous neuronal oscillations synchronize with auditory stimuli. Although invasive methods such as intracranial electrography are excellent tools for exploring this [2], such methods are not suitable for healthy subjects. Electroencephalograph (EEG) is an effective, non-invasive technique for investigating the neural mechanism behind auditory processing. Oscillations observed in the EEG signal are direct reflections of neural oscillations in the cortex [3]. However, the generated electrical fields are easily contaminated by external noise (e.g., eye movement, heartbeat) that occur during the transmission from the neural population to the top layer of the scalp through the brain tissue and skull. Such a non-invasive technique is often limited by the low signal-to-noise ratio (SNR) of the recording neural signal. Event-related potential (ERP) is often used to improve the SNR [4]. The neural response to a stimulus should be similar across all trials, and randomly distributed noise should be considered independent from the response. Thus, the noise can be decreased as a trial is repeated, and the meaningful response (the ERP) can be calculated by averaging the repeats. As such, well-designed trials need to be repeated sufficiently in order to use ERP to improve the SNR.

However, it may be impossible to repeatedly conduct trials in certain paradigms. In recent years, studies have extended the controlled experimental paradigm to a more natural setting [5], leading to investigations on how neural activity synchronizes with the acoustic or linguistic information of a continuous naturalistic speech stream (neural entrainment to speech) [6–9]. In neural entrainment research, stimuli are typically long segments (around 15 s to 120 s) from lectures or stories and presented to subjects only once to avoid a priming effect. The main problem of these studies is how to accurately estimate the temporal response functions (TRFs) of the neural system [10]. If we treat the neural system as a linear system, TRFs can help to linearly map speech stimulus representations (e.g., speech envelope, fine structure, spectrograms) to the neural response of the listener. TRFs can enable us to investigate how neural oscillations modulate the speech signal [11] or localize the brain regions involved with the speech processing by using the source reconstruction method [12]. A method of accurately estimating TRFs is necessary in these studies. As mentioned above, it is impossible to specify ERPs by averaging more trials in the neural entrainment studies. TRFs are often estimated from a single trial. It is difficult to reduce environmental noise when using the single-trial method. Furthermore, in long listening tasks, it is more difficult to detect unexpected noise caused by breathing or lack of attentiveness, which negatively impacts TRF estimation.

Assuming brain functions for speech processing are consistent across individuals, a similar neural response is expected for the same speech stimulus. In contrast, external noise, involuntary breathing, and attentiveness differ from individual to individual. Such noises can be suppressed by averaging the neural signal of the same stimuli for all subjects. In our

previous work [13], we have proven the effectiveness of this idea. However, this is suboptimal to apply due to the lack of a method to account for the subjects' differences in the latencies of the neural response to the stimulus, as well as differences in setup positions of the electrodes. Addressing these two problems well before averaging the neural activities across multiple subjects should result in a more accurate TRF estimation. Thus, in this study, we propose a temporal-spatial hyper-alignment method which uses a well-designed spatial filter to align the latency time at the temporal domain and finds the common neural response at the spatial domain for eliminating these effects. Then we constructed a neural entrainment model to study the neural coherence to speech.

This paper is organized as follows. Section II introduces the experimental design and describes the proposed hyper-aligned methods in detail. Our results are reported in Section III, and our conclusions are given in Section IV.

II. Materials and Methods

A. Participants

Twenty-two healthy Mandarin Chinese speakers (mean ± standard deviation age, 22 ± 2.4 years; nine men; right-handed) were recruited from Tianjin University and Tianjin University of Finance and Economics. The experiments were conducted in accordance with the Declaration of Helsinki [14] and approved by the local ethics committee. The subjects signed informed consent forms before the experiment and were paid for their participation afterward. All the subjects reported no history of hearing impairment or neurological disorders.

B. Stimuli and experimental procedure

Subjects undertook 24 non-repetitive trials; each trial was a short story (around 60 s) with a complete storyline, recorded by a male Chinese announcer in a soundproof room. All stimuli were mono speech with a 44.1 kHz sampling rate, and the stimulus amplitudes were normalized to have the same root mean square (RMS) intensity. The 24 trials were randomly presented to the subjects. All speech segments were also modified to truncate the silence gaps to less than 0.5 s [7].

The experiment was carried out in an electronically and magnetically shielded soundproof room. In the experiment, speech sounds were presented to subjects through Etymotic Research ER-2 insert earphones (Etymotic Research, Elk Grove Village, IL, USA) at a suitable volume (around 65 dB). During each trial, subjects were instructed to focus on a crosshair mark in the center of the screen to minimize blinking, head movements, and other bodily movements. There was a five-second interval between each trial, and the subjects were given a five-minute break every ten trials. After each story trial, subjects were asked immediately to answer multiple-choice questions about the content of the story to ensure that they focused on the auditory task. We embedded unique tones in some trials to draw more of the subjects' attention to the stimuli. Subjects were requested to detect the tones and indicate how many times they appeared after the trial. The EEG data corresponding to the embedded tones were removed in further analysis.

C. EEG data acquisition and pre-processing

The scalp EEG signal was recorded with a 128-channel Neuroscan Synamps system (Neuroscan, USA) at a sampling rate of 1000 Hz. The electrodes were placed according to the standard 10-5 system, and six channels were used for recording a vertical electrooculogram (VEOG), a horizontal electrooculogram (HEOG), and two mastoid signals. The impedance of each electrode was kept below 5

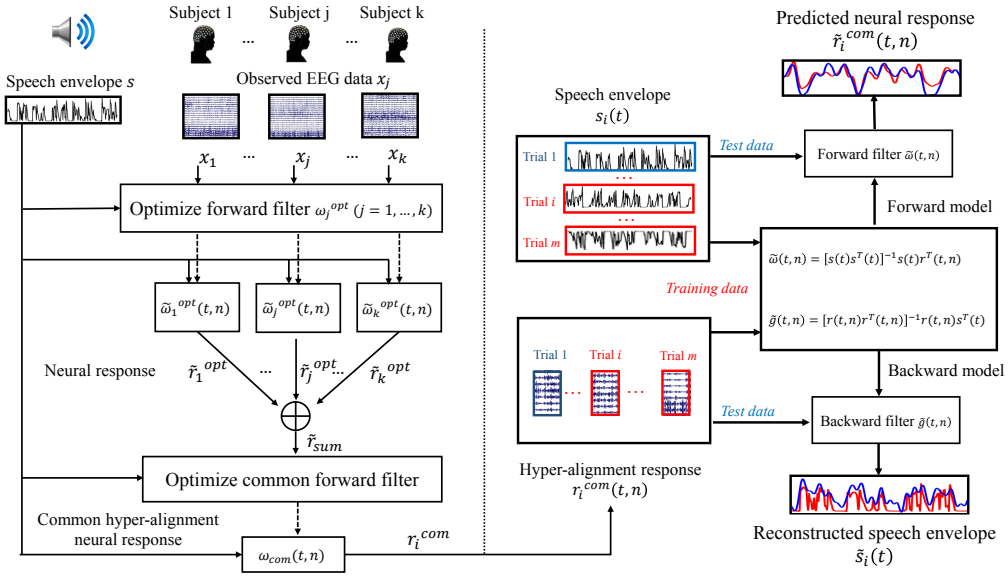

Fig. 1: Pre-processing and training procedures.

kΩ during data acquisition. Three subjects' data were discarded in the further analysis because they did not give a proper answer for the multiple-choice questions or the electrodes detached during the EEG data recording. The raw EEG data were pre-processed using the EEGLAB toolbox (https://sccn.ucsd.edu/eeglab/index.php) in MATLAB (MathWorks) [15]. This involved removing sinusoidal (i.e., line) noise and bad channels (i.e., low-frequency drifts, noisy channels, short-time bursts) and repairing the data segments. Then, the EEG data was bandpass filtered in the delta band (1–3 Hz), theta band (4–8 Hz), and then downsampled to 64 Hz [6, 16]. As previous research has shown that speech envelopes are more relevant to 1–8 Hz [10, 17], we focused on the delta and theta bands in this study.

The broadband temporal speech envelopes were obtained from Hilbert transforms. For the following modeling approach, the envelope was then decimated to the same sampling rate as EEG, enabling us to relate their dynamics to the EEG signals. During the experiment, we marked the time trigger for the EEG signal according to the stimuli onset and offset. In the offline analysis, the 24 data epochs (24 story trials) were extracted on the basis of the time trigger for each subject (19 subjects × 24 trials). We assume that all of the subjects use the same neural mechanism to process the stimulus speech so their TRFs are nearly the same. The averaged alignment data on all subjects is expected to reduce the noises which may be caused by breathing, inattentiveness, etc., through averaging processing. Before averaging, a spatial filter was designed to align the latency of neural responses for accurately extracting the common neural response across subjects.

D. Procedure of proposed method

Consider $x(t,n)$ is the observed EEG raw data of channel n at time index t and $r(t,n)$ is the stimuli related neural response. Thus, $x(t,n)$ can be expressed as:

$$x(t,n) = r(t,n) + rest(t,n), \tag{1}$$

where $rest(t,n)$ is the residual noise unrelated to the stimuli, including environmental noise and noise caused by breathing, inattentiveness, and unexpected events. As we assumed that different subjects' brains are functionally similar, the similar stimuli related neural responses $r_j(t,n)$ can be expected. And the residual noises $rest_j(t,n)$ differ from individual to individual. After averaging the observed EEG raw data $x(t,n)$ of the same stimuli for all subjects, such noise can be effectively suppressed [13]. However, due to the subjects' differences in the latencies of the neural response to the stimulus. Before averaging, the latency of neural responses for the subjects should be aligned.

Given two different subjects' observed EEG raw data $x_1(t,n)$ and $x_2(t,n)$, our proposed method produces spatial filters (transform matrices) p_1 and p_2 to align the latency of neural responses. According to our hypothesis, after we optimize the latency of neural responses, the transform result $\widetilde{r}_1(t,n) = p_1 x_1(t,n)$ and $\widetilde{r}_2(t,n) = p_2 x_2(t,n)$ should be similar, and both of them have the highest correlation with the stimuli. More importantly, corresponding columns from $\widetilde{r}_1(t,n)$ and $\widetilde{r}_2(t,n)$ are also maximally correlated with each other so that they can be averaged and not be affected by the different electrodes' location. Therefore, our proposed method can also align the different position of electrode across subjects.

Accordingly, Our purpose is to make the $x(t,n)$ most relevant to the real response $r(t,n)$, which also means to eliminate noises unrelated to the stimuli to obtain the true neural response $r(t,n)$ from the observed $x(t,n)$. Here, we try to explain how to use spatial filters to reduce the unrelated residual noise. Spatial filters were constructed by optimizing the importance (weights) for each electrode. Assuming a spatial filter p exists, most of the noise $rest(t,n)$ can be denoised by multiplying $x(t,n)$ with p; in other words, the SNR of $x(t,n)$ is maximized by the spatial filter p,

$$\arg\max_p \frac{E\{[p^T r(t,n)]^2\}}{E\{[p^T rest(t,n)]^2\}} = \arg\max_p \frac{p^T R_{rr} p}{p^T R_{restrest} p}. \tag{2}$$

The neural response $r(t,n)$ and $rest(t,n)$ are independent. Therefore, (2) is equivalent to the new equation which maximizes the ratio of neural response to the observation signal [18],

$$\arg\max_p \frac{p^T R_{rr} p}{p^T (R_{rr} + R_{restrest}) p} = \arg\max_p \frac{p^T R_{rr} p}{p^T R_{xx} p}. \tag{3}$$

In other words, multiplying p (the solution for p will be explained in a later subsection) is equivalent to minimizing the mean-squared error (MSE) between the true neural response $r(t,n)$ and the observed $x(t,n)$,

$$\widetilde{r}(t,n) = \arg\min_r E\{\sum_t [x(t,n) - r(t,n)]^2\}. \tag{4}$$

Here, we need to estimate the unknown $r(t,n)$. As mentioned above, brain functions are considered to be a linear time-invariant (LTI) system where the output (neural response) of the system is the convolution of the input and a TRF of the brain. The TRF can be considered a filter that linearly transfers the continuous speech envelope to the dynamic neural response. The TRF of the channel n is a function of $\omega(t,n)$ of time t and the output of the neural system is $r(t,n)$ for the same channel n. For an input speech stimulus $s(t)$, the output can be described as:

$$r(t,n) = \sum_\tau \omega(\tau,n) s(t-\tau), \tag{5}$$

as the latency of the neural response differs between subjects. The optimal latency can be used to obtain the best solution for $\widetilde{\omega}(t,n)$. Therefore, (4) can be changed to

$$\begin{aligned} \widetilde{\omega}_j^{opt}(t,n) = \arg\min_\omega E\{\sum_t [x_j(t,n) - \sum_\tau \omega(\tau,n) \\ s(t-\tau-lag)]^2\} \ (0 \le lag \le 400ms), \end{aligned} \tag{6}$$

$\widetilde{\omega}_j^{opt}(t,n)$ is the optimal TRF function with a response latency optimized for subject j. Thus, the neural response $\widetilde{r}_j^{opt}(t,n)$, n can be used to approximate the true response of subject j.

Next, we summate the neural response $\widetilde{r}_j^{opt}(t,n)$ of all subjects. The spatial transformation eliminates the effects of electrode position in the device setup so that the subjects' data can be summated or averaged. Thus, the summation enables us to identify the common neural response and decrease most noises caused by individual differences such as breathing or inattentiveness.

$$\widetilde{r}_{sum}(t,n) = r_{com}(t,n) + rest_{noise}(t,n), \tag{7}$$

where the $\widetilde{r}_{sum}(t,n)$ is the summation of the subjects' neural response. $r_{com}(t,n)$ is the common neural response related to the stimuli across all subjects, and $rest_{noise}(t,n)$ is the residual noise from the previous estimation procedure. As a result, Eq. (8) has a similar structure to (1). We can apply another spatial filter p_{com} to reduce $rest_{noise}(t,n)$. Then, the common neural response $r_{com}(t,n)$, which maximizes the ratio to $\widetilde{r}_{sum}(t,n)$, can be obtained by the same procedure in (2). We reduced both the effect of latency in the temporal domain and electrode position in the spatial domain. We call our proposed approach the temporal-spatial hyper-alignment method. The pre-processing procedure of our proposed method is shown in the left-hand side of Fig. 1.

E. Training procedure for neural entrainment modeling

In this study, we used an mTRF toolbox (https://github.com/mickcrosse/mTRF-Toolbox) to linearly map the speech envelope and the neural response [19]. The main principle is to treat the brain as an LTI system. Then, the forward model $\omega(t,n)$ is defined to predict the neural response $r(t,n)$ for an input speech stimulus $s(t)$ in (5). Here, the solution of $\omega(t,n)$ is:

$$\widetilde{\omega}(t,n) = [s(t)s^T(t)]^{-1}s(t)r^T(t,n). \tag{8}$$

In a hypothetical LTI system, a backward approach can be modeled using a decoder $g(t,n)$, which is the inverse function of $\omega(t,n)$. Thus, the input speech stimulus $s(t)$ can be reconstructed by filtering the neural response $r(t,n)$ using the decoder function $g(t,n)$. This can be expressed as:

$$s(t) = \sum_n \sum_\tau g(\tau,n)r(t-\tau,n), \tag{9}$$

where $\widetilde{s}(t)$ is the reconstructed speech stimuli. The optimal decoder $g(t,n)$ is acquired by minimizing the MSE between the original and reconstructed speech stimuli. The solution of $g(t,n)$ is:

$$\widetilde{g}(t,n) = [r(t,n)r^T(t,n)]^{-1}r(t,n)s^T(t). \tag{10}$$

To evaluate our method, we used the temporal-spatial hyper-aligned EEG to predict the neural response and reconstruct the speech envelope by the forward and backward model, respectively. Then, we compared the neural response prediction and speech envelope reconstruction accuracy of the proposed hyper-alignment method with those of the single-trial based method. From pre-processing the hyper-alignment neural response $r_i^{com}(t,n)$ was obtained, where i is the number of trials

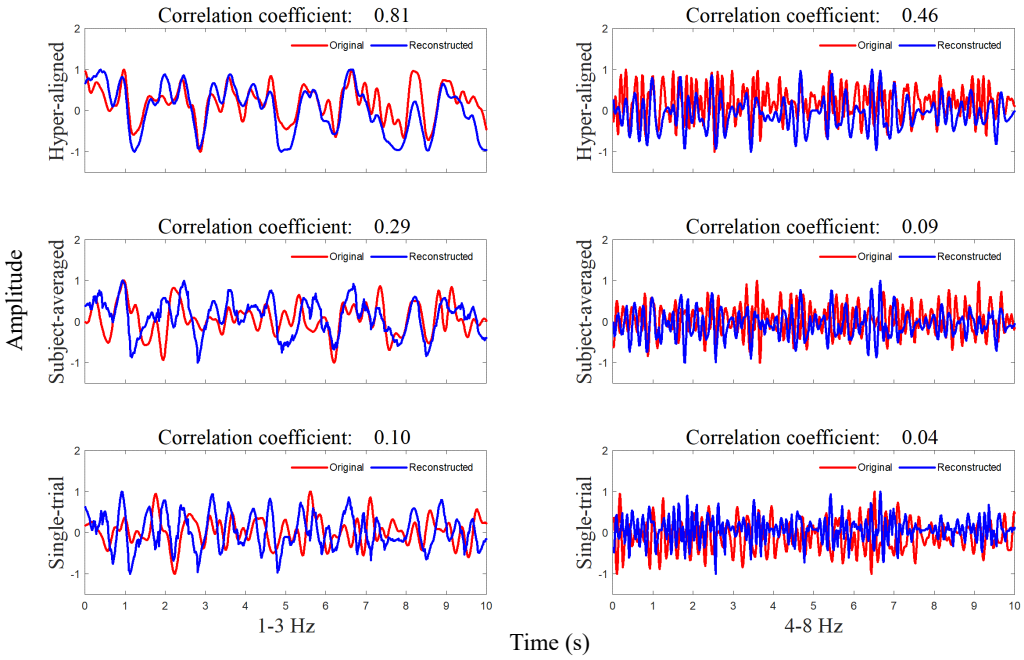

Fig. 2: Examples of neural responses predicted by different methods.

$(1 \leq i \leq m)$. For all training processes, we used a leave-one-out cross-validation procedure, where 23 trials were used for training, and the remaining one trial was used for testing in each fold. Because the parameters of $\widetilde{\omega}(t,n)$ and $\widetilde{g}(t,n)$ were different in each trial, we used the averaged parameters of the forward filter $\widetilde{\omega}(t,n)$ and backward filter $\widetilde{g}(t,n)$ trained on the other 23 trials [20].

In the single-trial based method, the forward and backward model was trained based on the subject's neural data. Since each subject took part in 24 trials, the procedure was repeated 24 times for each subject. Our proposed method was trained based on the averaged subject hyper-alignment data, which was repeated for 24 iterations for the hyper-alignment data. Our training procedure is shown in the right-hand side of Fig. 1.

F. Estimation for spatial filter p

As stated in subsection D, R_{rr} can be easily estimated from the TRFs model. R_{xx} can be calculated from the recorded raw EEG data. Only the spatial filter p is unknown in (3). According to previous research [18, 21], to obtain the maximum SNR, the stationary points of (3) must satisfy

$$R_{rr}p = \lambda R_{xx}p, \tag{11}$$

which defines a generalized eigenvalue problem (GEVP) for the matrix pencil (R_{rr}, R_{xx}). All λ and p that can be substituted into (11) are denoted as the generalized eigenvalues and eigenvectors [18]. When multiplying the spatial filter p^T on both sides of (11), we can get

$$p^T R_{rr}p = \lambda p^T R_{xx}p, \tag{12}$$

where

$$\lambda = \frac{p^T R_{rr}p}{p^T R_{xx}p}, \tag{13}$$

which implies that λ is proportional to the output SNR of the (3). Therefore, in order to maximize the SNR, spatial filter p should be set to be the generalized eigenvector which corresponds to the maximum eigenvalues λ.

III. Results

A. Behavioral results

To verify the situation of subjects during the experiment, we asked the subjects to answer multiple-choice questions about the story presented in the listening task after each trial. The accuracy of the answers was 88.25 ± 4.62%, indicating that most of the subjects concentrated on the listening task during the experiment. We removed three subjects' data since their answers were not sufficiently accurate.

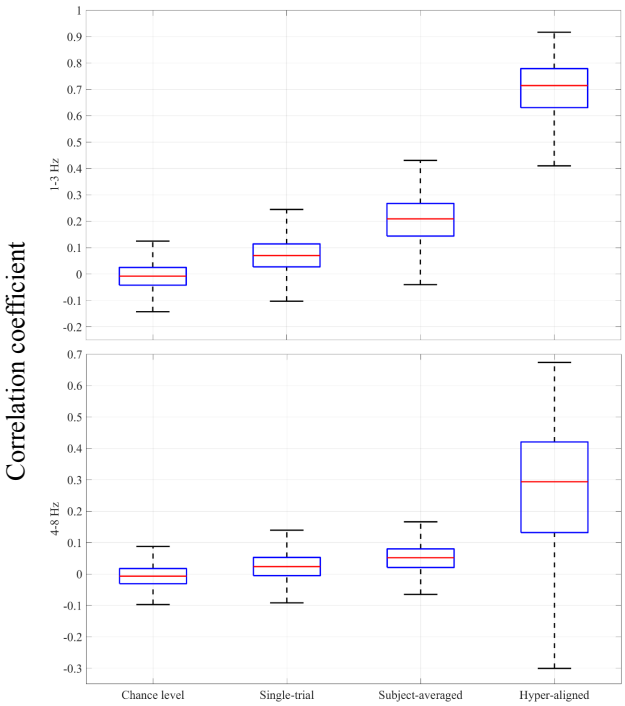

Fig. 3: Comparison of neural response prediction accuracies between proposed hyper-aligned method and other two methods.

B. Neural response prediction results

In the forward process, we used the forward TRF to predict the 122 channels' neural activity response to the speech envelope. The prediction accuracy was evaluated by measuring the Pearson correlation coefficient between the predicted neural signals and the original ones (for our proposed method, the original signal is the averaged one of all subjects). Fig. 2 shows examples of the predicted neural responses obtained in our study. The correlation coefficients of our proposed method are 0.81 and 0.46 in the delta band (1–3 Hz) and theta band (4–8Hz) respectively, which is significantly higher than that of other methods. Fig. 3 shows the comparisons of the averaged correlation of 122 channels for the proposed method and the other two methods in the delta and theta bands. To quantitatively compare the three methods, the correlation coefficient was firstly transformed into a z value by Fisher's z transformation to satisfy a normal distribution [22]. Then, an analysis-of-variance (ANOVA) of the z values with factors of frequency (different frequency bands) and the method (proposed method and the other two methods) revealed a significant effect on both frequencies ($F = 49847, p < 0.001$) and reconstruction methods ($F = 90479, p < 0.001$). The results of ANOVA demonstrated that the prediction accuracy of our proposed method is higher than that of the other two methods. We used a permutation test to compare the predicted accuracy and the chance level and found that our prediction

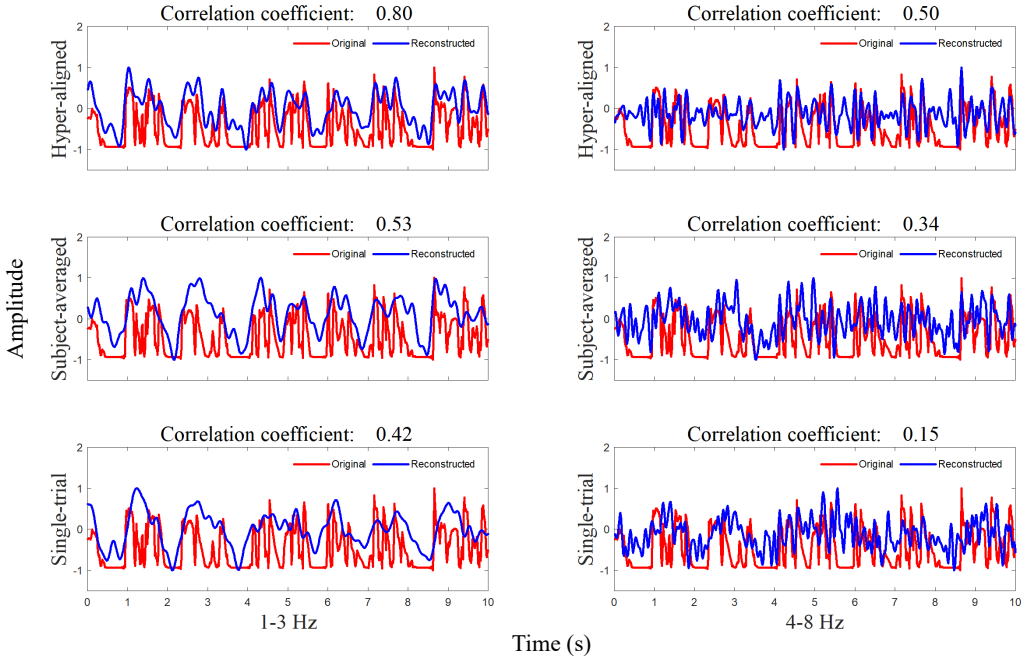

Fig. 4: Examples of original speech envelopes and the reconstructed ones using different methods.

value is 288 times larger than that of the chance level.

C. Speech envelope reconstruction results

In the backward process, the speech envelope was reconstructed from the neural response. The reconstruction accuracy is represented by the correlation between the reconstructed speech envelope and the original one. Fig. 4 shows examples of the reconstructed envelopes obtained in our study. The correlation coefficients for the reconstructed speech envelope of our proposed method are 0.80 and 0.50 in the delta band (1–3 Hz) and theta band (4–8Hz) respectively, which is also significantly higher than that of other methods. Fig. 5 shows the comparisons of reconstruction accuracies for the proposed and the other two methods in delta and theta bands. The reconstruction accuracy was significantly higher than the chance level in the delta (1–3 Hz) and theta (4–8 Hz) bands (Fig. 5). Similar to the previous section, the values of correlation coefficients were also converted to z values using Fisher's z transformation to satisfy a normal distribution. An ANOVA of the z values with the main factors of frequency and reconstruction method revealed a significant effect caused by the reconstruction methods ($F = 830.39, p < 0.001$), indicating that the accuracy of speech envelope reconstruction of our hyper-aligned method is higher than that of the other two methods in the two frequency bands ($F = 887.89, p < 0.001$).

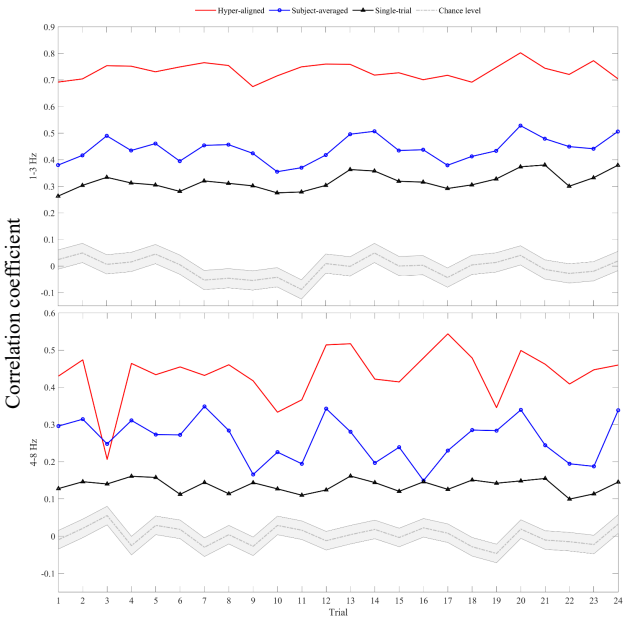

Fig. 5: Comparison of speech envelope reconstruction accuracies in each trial between proposed hyper-aligned method and other two methods.

D. Robustness of proposed method

To verify whether or not the proposed method is data dependent, we used an open dataset for testing (Dryad, https://datadryad.org/stash/dataset/doi:10.5061/dryad.070jc) which was used in previous studies [6, 8]. Our proposed method obtained an average accuracy of 0.18 and 0.31 for predicting the neural response in 1–15 Hz and 1–4 Hz, respectively, which are much higher than the previous average accuracy of 0.06 and 0.04 [6, 8]. Since our proposed method

eliminates individual differences, it is difficult to use the TRF model based on our method to investigate individual properties. Rather, our method can obtain a more general neural entrainment TRF model with high accuracy.

IV. Conclusions

In this study, we assumed that brain functions for speech processing are consistent across individuals. Based on this assumption, averaging responses over multiple subjects is expected to be an efficient way to improve the accuracy of TRF estimation. However, individual differences in brain anatomy and device setup in the experiment are difficult to account for in the TRF modeling approach, so the subject-averaged method is a sub-optimal solution. Inspired by the multiway canonical correlation analysis [23] and data-driven stimulus-related neural activity selection method [18], we proposed a multi-subject hyper-alignment method to reduce those individual differences. Compared with the single-trial based and subject-averaged methods, our hyper-aligned method obtained the highest accuracy for the predicted neural response and reconstructed speech envelope. The mean correlation between the reconstructed speech envelope and the original increased to about 0.73 in the delta band (Fig. 5). Compared with the single-trial based envelope reconstruction accuracy of 0.32, the error reduction rate was around 60%. The prediction accuracy was also improved in the theta band. The robustness and universality of our proposed method were also verified by using a different EEG dataset.

In this study, we used a coarse speech feature – the speech envelope. In future work, we intend to integrate more detailed features to investigate the neural mechanisms of speech processing.

Acknowledgment

This study is supported in part by JSPS KAKENHI Grant (20K11883), and in part by National Natural Science Foundation of China (No.61876126). The authors thank Jinfeng Huang for useful discussions and help on the EEG data collection.

References

[1] G. Zhang, Y. Si, and J. Dang, "Revealing the Dynamic Brain Connectivity from Perception of Speech Sound to Semantic Processing by EEG," *Neuroscience,* vol. 415, pp. 70–76, 2019.

[2] G. K. Anumanchipalli, J. Chartier, and E. F. Chang, "Speech synthesis from neural decoding of spoken sentences," *Nature,* vol. 568, no. 7753, pp. 493–498, 2019.

[3] M. X. Cohen, *Analyzing neural time series data: theory and practice.* MIT press, 2014.

[4] T. C. Handy, *Event-related potentials: A methods handbook.* MIT press, 2005.

[5] J. Brennan, "Naturalistic sentence comprehension in the brain," *Lang. Linguist. Compass,* vol. 10, no. 7, pp. 299–313, 2016.

[6] G. M. Di Liberto, J. A. O'Sullivan, and E. C. Lalor, "Low-frequency cortical entrainment to speech reflects phoneme-level processing," *Curr. Biol.,* vol. 25, no. 19, pp. 2457–2465, 2015.

[7] C. Brodbeck, A. Presacco, and J. Z. Simon, "Neural source dynamics of brain responses to continuous stimuli: Speech processing from acoustics to comprehension," *Neuroimage,* vol. 172, pp. 162–174, 2018.

[8] M. P. Broderick, A. J. Anderson, G. M. Di Liberto, M. J. Crosse, and E. C. Lalor, "Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech," *Curr. Biol.,* vol. 28, no. 5, pp. 803–809, 2018.

[9] Pan, X. Y., Zou, J. J., Jin, P. Q., and Ding, N, "The neural encoding of continuous speech-recent advances in EEG and MEG studies," *Sheng li xue bao Acta Physiol. Sin.,* vol. 71, no. 6, pp. 935–945, 2019.

[10] N. Ding and J. Z. Simon, "Neural coding of continuous speech in auditory cortex during monaural and dichotic listening," *J. Neurophysiol.,* vol. 107, no. 1, pp. 78–89, 2012.

[11] A. Kösem and V. Van Wassenhove, "Distinct contributions of low- and high-frequency neural oscillations to speech comprehension," *Lang. Cogn. Neurosci.,* vol. 32, no. 5, pp. 536–544, 2017.

[12] P. Das, C. Brodbeck, J. Z. Simon, and B. Babadi, "Neuro-current response functions: A unified approach to MEG source analysis under the continuous stimuli paradigm," *Neuroimage,* vol. 211, p. 116528, 2020.

[13] Di Zhou, Gaoyan Zhang, Jianwu Dang, Shuang Wu, and Zhuo Zhang, "Neural Entrainment to Natural Speech Envelope Based on Subject Aligned EEG Signals," in *Interspeech, 2020*.

[14] G. A. of the W. M. Association, "World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects," *J. Am. Coll. Dent.,* vol. 81, no. 3, p. 14, 2014.

[15] A. Delorme and S. Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *J. Neurosci. Methods,* vol. 134, no. 1, pp. 9–21, 2004.

[16] O. Etard and T. Reichenbach, "Neural Speech Tracking in the Theta and in the Delta Frequency Band Differentially Encode Clarity and Comprehension of Speech in Noise," *J. Neurosci.,* vol. 39, no. 29, p. 5750, 2019.

[17] J. Zou et al., "Auditory and language contributions to neural encoding of speech features in noisy environments," *Neuroimage,* vol. 192, pp. 66–75, 2019.

[18] N. Das, J. Vanthornhout, T. Francart, and A. Bertrand, "Stimulus-aware spatial filtering for single-trial neural response and temporal response function estimation in high-density EEG with applications in auditory research," *Neuroimage,* vol. 204, p. 116211, 2020.

[19] M. J. Crosse, G. M. Di Liberto, A. Bednar, and E. C. Lalor, "The multivariate temporal response function (mTRF) toolbox: a MATLAB toolbox for relating neural signals to continuous stimuli," *Front. Hum. Neurosci.,* vol. 10, p. 604, 2016.

[20] J. A. O'Sullivan et al., "Attentional selection in a cocktail party environment can be decoded from single-trial EEG," *Cereb. cortex,* vol. 25, no. 7, pp. 1697–1706, 2015.

[21] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE assp Mag.,* vol. 5, no. 2, pp. 4–24, 1988.

[22] D. M. Corey, W. P. Dunlap, and M. J. Burke, "Averaging correlations: Expected values and bias in combined Pearson rs and Fisher's z transformations," *J. Gen. Psychol.,* vol. 125, no. 3, pp. 245–261, 1998.

[23] A. de Cheveigné et al., "Multiway canonical correlation analysis of brain data," *Neuroimage,* vol. 186, pp. 728–740, 2019.